DOCUMENT RESUME

ED 147 711 CG 012 047

AUTHOR Nelson, Gary L.
TITLE Methodology in Token Economy Evaluation: A Review and Critique.
PUB DATE [77]
NOTE 17p.; Best copy available

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Behavior Change; *Contingency Management; *Evaluation Methods; *Institutional Environment; Literature Reviews; Patients (Persons); Psychiatric Hospitals; *Psychotherapy; Social Psychology; *Therapeutic Environment
IDENTIFIERS *Token Economy

ABSTRACT

Major problem areas in evaluating token economies for psychiatric inpatients are identified and discussed. Three categories of methodological deficiencies frequently encountered in the token economy literature are described: (1) inadequate specification of treatment procedures; (2) insufficient or inadequately described dependent measures; and (3) use of an experimental design which does not permit demonstration of a cause-effect relationship between treatment and behavior change. Guidelines for evaluating methodology in each category are presented and then applied in a critical review of token economy studies. It was concluded that no study provided clear evidence of a cause-effect relationship between clearly specified treatment procedures and changes on objective dependent measures adequate for treatment evaluation. As a result, the status of the token economy as a treatment modality for psychiatric inpatients is unclear. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147711

Methodology in Token Economy Evaluation: A Review and Critique

Gary L. Nelson
West Virginia University

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gary L. Nelson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

CG 012 047

Abstract

Major problem areas in evaluating token economies for psychiatric inpatients are identified and discussed. Three categories of methodological deficiencies frequently encountered in the token economy literature are described: (a) inadequate specification of treatment procedures, (b) insufficient or inadequately described dependent measures, and (c) use of an experimental design which does not permit demonstration of a cause-effect relationship between treatment and behavior change. Guidelines for evaluating methodology in each category are presented and then applied in a critical review of token economy studies. No study provided clear evidence of a cause-effect relationship between clearly specified treatment procedures and changes on objective dependent measures adequate for treatment evaluation. As a result, the status of the token economy as a treatment modality for psychiatric inpatients is unclear.

GG012047

Methodology in Token Economy Evaluation: A Review and Critique

Gary L. Nelson
West Virginia University

Token economy programs, as direct applications of laboratory-based operant methodology in clinical settings, typically involve extensive recording of response frequencies and thus offer the potential for systematic program evaluation. However, although the effects of token reinforcement on specific response frequencies within a token economy have been demonstrated (Ayllon & Azrin, 1965), an experimental demonstration of overall program effectiveness has been decidedly more difficult to achieve (Gripp & Magaro, 1974). Numerous token economy evaluations have been reported since Ayllon and Azrin's (1965) description of the effectiveness of token reinforcement with psychotic females. However, the inherent complexity of token economies as social-psychological systems (Kazdin, 1977; Krasner & Atthowe, 1971) as well as practical limitations imposed on research in applied settings, has resulted in most studies yielding findings which must be considered tentative rather than conclusive. Some of the major problems and areas of insufficient investigation which become evident upon careful analysis of the literature on token economies for psychiatric inpatients are described below.

1. Few attempts have been made to facilitate or evaluate generalization of behavior change beyond the hospital setting. Kazdin and Bootzin (1972) point out that few attempts have been made to program generalization to the non-hospital environment either by fading or thinning controlling stimuli before the patient leaves the treatment program, or by programming important antecedent and consequent stimuli into the post-treatment environment through such procedures as training relatives to use contingent reinforcement, or teaching patients to use self-reinforcement techniques. In addition to the relative lack of attempts to facilitate generalization, there is a scarcity of studies investigating generalization or maintenance of reported behavior changes once the patient has left the treatment program.

2. Few investigators have tried to program token reinforcement of complex behaviors such as language and social behavior. Though some investigators (e.g., Bennett & Maley, 1973) have attempted token reinforcement of complex social behaviors, token economies are typically directed at increasing basic skills such as self-care and work performance. Kazdin and Bootzin (1972) suggest that any general methodology of behavior change must address itself to more complex social behaviors, in spite of practical difficulties in measurement and contingency management.

3. There has been little attempt to identify and describe unresponsive subjects, or the reasons for their lack of response. Several

---

I thank John D. Cone for his valuable suggestions during the preparation of this paper.

reports (Allen & Margaro, 1971; Carlson, Hersen, & Eisler, 1972; Kazdin & Bootzin, 1972) discuss the fact that responsiveness to token reinforcement procedures varies widely between subjects, and many studies report that the behavior of some subjects was not altered. Adequate investigations have not been carried out to determine if individual unresponsiveness is due to insufficient programming, or to important individual difference variables that interact with token procedures.

4. Operative variables in token economies have not been adequately identified and analyzed. Gripp and Magaro (1974) describe token economies as complex social-psychological systems involving numerous variables which operate to modify patient behavior. They maintain that recognition and control of the many variables operating in token economies are necessary to evaluate the effects of specific operant procedures. Few attempts have been made to identify and determine the relative importance of the various components of the token economy treatment package. In fact, specification of the exact procedures used to modify patients' behavior is frequently omitted from reports describing token programs.

5. There is a scarcity of investigations which compare token economies with other clearly specified treatment approaches. Several authors (Agras, 1972; Gripp & Magaro, 1974) have emphasized that the ultimate practicality of any treatment approach depends in part on its effectiveness relative to other treatment methods. At the time of this writing, only two studies (Marks, Sonoda, & Schalock, 1968; Tenbrunsel, 1969) were found which attempted to compare a token economy with another specific form of treatment.

6. A "uniformity myth" (Kiesler, 1971) has developed regarding the token economy as a treatment package. Although discussions of a uniformity myth have not been evident in existing literature, a careful analysis of token economy research suggests that the token economy is often regarded as a uniform treatment approach which is the same across settings. Actually, most token programs represent independent attempts to apply operant principles in group settings such as hospital wards. The target behaviors, back-up reinforcers, amount and type of staff, procedures to initiate behaviors, and other variables may differ greatly across settings. The fact that various programs use a medium of exchange (i.e., tokens) which subjects use to obtain desired reinforcers, does not make the programs uniform or even similar in regard to the actual procedures used to initiate and maintain behavior change. Since token economies do differ widely in the specific procedures used, evidence for the effectiveness of a particular token economy cannot be appropriately cited as proof of the effectiveness of token economies in general. Considering the variation from one token economy to the next, Gripp and Magaro's (1974) emphasis on establishing the relative importance of various operative variables seems less relevant. To isolate and evaluate components in a single token economy may not reveal much at all about how the components operate in another token economy, especially since each program deals with a different group of patients. Even within a single program, the treatment components may differ in their relative effectiveness with different classes of behavior. The variability between programs and the probability of differential effectiveness across subjects and target behaviors highlights the importance of each token program report providing clear description of the specific treatment procedures used.

7. Token economy evaluation studies typically suffer from important methodological weaknesses. Token economies are complex social-psychological systems (Krasner & Atthowe, 1971; Gripp & Magaro, 1974), token economy evaluation is still a relatively new pursuit, and research in applied settings is often hampered by institutional restrictions (Gripp & Magaro, 1974; Kazdin & Bootzin, 1972). These factors in combination provide ample reason for the current lack of conclusive evidence regarding the effectiveness of token economy programs.

The methodological deficiencies which occur most frequently in token economy studies can be divided into three main categories: (a) inadequate specification of treatment procedures, (b) insufficient or inadequately specified dependent measures, and (c) use of an experimental design which does not allow demonstration of a cause-effect relationship between treatment procedures and reported behavior changes. The following section will discuss methodological issues in more detail.

Methodology in Token Economy Research

Specification of Treatment Procedures

In describing guidelines for research in behavior modification, Paul (1969) emphasized the advantages of obtaining homogeneity of treatment techniques across therapists, and suggested written manuals and demonstrations as means of gaining treatment uniformity. He pointed out that without constant monitoring, replication of independent variables would not be possible and negative findings could be the result of misapplied procedures. Paul's suggestions regarding treatment standardization, however, may represent an ideal that may be particularly difficult to achieve in token economy research. In a well-designed token program target behaviors are determined by an actual assessment of the behavioral excesses and deficits present in the client group involved, and back-up reinforcers are selected on the basis of preferences observed in that client group. Thus, two very important aspects of program design (i.e., target behaviors and back-up reinforcers) may not be appropriate for standardization. A standardized procedure for selecting target behaviors and back-up reinforcers could be developed, but the selection procedure itself would constitute a relatively minor aspect of the treatment package. Other components of the treatment package such as time and place for reinforcement delivery, prompting techniques (posted rules, instructions, or time schedules; verbalized reminders; modeling of desired behavior; etc.), and the number and type of staff administering the program, may also be very difficult to standardize for application across settings. Given the complexity of token economies, clear specification of the procedures used in each program, rather than strict standardization of procedures across programs, may be the appropriate route to establishing a useful body of knowledge concerning token economy effectiveness.

Along with treatment standardization, detailed specification of treatment procedures used in token programs is also subject to practical restriction. A token economy is typically carried out over an extended period of time, and procedures are frequently modified to fit changes in patient behavior or other changes which occur in the treatment milieu. To describe all procedures in detail, besides being extremely time consuming, would probably result in a lengthy document of little interest to most readers, and subject to extensive editing by prospective publishers. A

logical compromise would be to provide a general description of the types of procedures used throughout a program, but include precise, detailed specification of any procedures which are claimed to be causally-related to changes in rates of specific target behaviors.

Selection of Dependent Measures

Paul (1969) stated that the minimum requirement for measuring effectiveness in an experiment on behavior modification is that change in the client's distressing behaviors from pretreatment to post-treatment be assessed external to treatment proper. Such a goal may be based on an analysis of individual or group treatment situations in which clients enter treatment to relieve themselves of distressing behavior patterns exhibited outside of treatment. In a token economy, many patients may not leave the treatment environment, and thus not be subject to assessment outside of the treatment setting. In addition, those who do leave may move to remote environments in which data regarding post-treatment behavior rates would be extremely difficult to obtain. Discharge and return rates have been used as global measures of improvement in individual functioning, and as indicators of generalized behavior change outside of the treatment setting (Atthowe & Krasner, 1968; Birky, Chambliss & Wasden, 1971; Gorham, Green, Caldwell & Bartlett, 1970; Heap, Boblitt, Moore & Hord, 1970; Lloyd & Abel, 1970; McReynolds & Coleman, 1972; Shean & Ziedberg, 1971; Steffy, Hart, Torney & Marlett, 1969). However, many investigators emphasize the inadequacy of discharge and return rates as measures of treatment effectiveness (Ayllon & Azrin, 1968; Birky, et al., 1971; Gripp & Magaro, 1974; Maley, Feldman & Ruskin, 1973). Such indices depend largely on a variety of factors other than patient behavior, such as the philosophy of staff making discharge decisions, the adequacy of community placement services, and the existence of after-care programs.

Kiesler (1971) states that there are no best measures that can be recommended for evaluating the outcome of behavior modification treatment. Rather, there are as many measures as may be required to assess the varied changes hypothesized by different theories of treatment. Since a token economy is a direct application of operant methodology in a group system, it would appear that the primary and necessary dependent measure would be rates of specific behaviors for which contingent reinforcement or punishment is applied.

Over and above reported changes in specific target behaviors, it would seem necessary to investigate the more global therapeutic effect of the program on individual participants. Since a token economy is supposedly a therapeutic program, it would appear appropriate to demonstrate not only that changes in basic self-care or work behaviors have occurred, but that the patient is in some way better off as a result. Some assessment of individual level of functioning, perhaps investigating such dimensions as mood, sociability, activity level, or degree of psychotic symptomatology, would therefore appear to be a required part of any token economy evaluation.

A third important domain for assessment involves more general effects of treatment on the overall ward environment. Although changes in rates of specific target behaviors and in levels of individual functioning guarantee some change in general behavior on the ward, it would appear important that certain dimensions of general ward behavior be investigated independently. Levels of general activity, appropriate social interaction, participation

in recreational activities, and involvement in work tasks, as well as frequencies of aggressive, disruptive, or bizarre behaviors, provide important indices of qualitative changes in the ward environment, and evaluation of any ward-wide treatment approach would seem incomplete without assessment of such factors.

The three domains for measurement discussed above (i.e., specific target behaviors, global individual functioning and general ward behavior) are suggested herein as basic measurement areas for assessing token economy effectiveness. Although the ultimate value of any therapeutic technique depends considerably on the durability of the positive changes produced, thorough assessment of therapeutic changes occurring within the treatment environment is a necessary and basic step before meaningful investigation of the durability of those changes following treatment can be conducted. Therefore, for purposes of this review, inclusion of procedures to measure change in the three domains specified is considered an adequate, though clearly not optimal, approach to token economy evaluation.

Experimental Design

Gripp and Magaro (1974) accurately point out that token economy research has involved a mixed bag of experimental designs typically dictated by extra-program considerations such as administrative demands. Campbell and Stanley (1963) classified experimental designs in education and psychology into three categories: (a) pre-experimental designs, (b) quasi-experimental designs, and (c) true experimental designs. Designs from all three categories have been used in token economy research. Pre-experimental designs used have included single-group studies with pre-post or repeated measures (McReynolds & Coleman, 1972; Steffy, et al., 1969) and static-group comparisons (Birky et al., 1971). Quasi-experimental designs have included time-series experiments (Atthowe & Krasner, 1968; Gorham, et al., 1970), non-equivalent control group studies (Heap et al., 1970; Gripp & Magaro, 1971; Shean & Ziedberg, 1971), and equivalent time-samples designs (Ayllon & Azrin, 1965; Lloyd & Garlington, 1968; Winkler, 1970). Those studies which have involved random assignment of subjects to experimental and control groups (Henderson & Scoles, 1970; Maley et al., 1973; Schaefer & Martin, 1966) are categorized as true experimental designs. Of the above designs, only the equivalent time-samples experiment and the control group studies with random assignment are generally considered adequate to establish a cause-effect relationship between independent and dependent variables (Campbell & Stanley, 1963; Paul, 1969).

In token economy research, the equivalent time-samples design has typically been used in the form of an ABA own-control design where A refers to a baseline period and B refers to the treatment (c.f. Ayllon & Azrin, 1965). In an ABA design, a causal relationship between treatment and behavior change is demonstrated if a subject's behavior improves when the treatment is presented and declines when the treatment is withdrawn. This design may have serious shortcomings in token economy research, as pointed out by Kazdin and Bootzin (1972). An ABA design is suitable only if the behavior studied is transient and reversible, and may be inappropriate when dealing with behaviors (such as grooming) which may easily come under control of nonprogrammed social reinforcers, and thus not decrease substantially under withdrawal conditions. In addition, the desirability of invoking withdrawal conditions on behavioral improvements gained in chronic mental patients is questionable, and is not likely to be approved by non-research-

oriented hospital staff. ABA designs have been employed to evaluate changes in discrete target behaviors, but have not been used to evaluate a total token economy program.

Control group designs with random assignment have been used infrequently in token economy research. These designs allow demonstration of a causal relationship between treatment and behavior change. However, adequate control group designs are extremely difficult to achieve in most applied settings because practical considerations typically make random assignment to groups (an essential feature of these designs) difficult, if not impossible (Kazdin & Bootzin, 1972).

Token economy research has thus been faced with the problem that experimental designs powerful enough to demonstrate a cause-effect relationship are either of questionable appropriateness (i.e., ABA designs) or extremely impractical (i.e., randomized control group designs). Much of token economy research has involved designs which can be classified as non-factorial single-group designs (Paul, 1969), and thought of essentially as combined case studies with either pretest-posttest or time-series measurement. Designs such as these do not control for possible effects of simultaneous extra-treatment events or maturation processes, and may also easily be confounded by testing effects or instrument errors. No cause-effect relationships can be established by these designs.

Kazdin and Bootzin (1972) suggest the multiple-baseline design (Baer, Wolf & Risley, 1968) or the multi-element baseline design (Sidman, 1960), as a possible approach to token economy evaluation. Multiple-baseline designs involve collection of data across behaviors, subjects, or situations. When data are collected across behaviors, two or more behaviors of a subject (or subjects) are monitored continuously. When behavior rates have stabilized, the experimental variable is introduced to one of the behaviors. Ideally, change occurs in the initial behavior while the other behaviors remain stable. The process is then repeated for each of the other behaviors. Extraneous variables are ruled out as alternative hypotheses if each behavior remains stable until the experimental variable is introduced and then changes in the direction predicted. Kazdin and Bootzin point out that the multiple-baseline design is particularly well suited to situations in which behavioral reversal is unanticipated or undesirable. However, they emphasize the inappropriateness of this design when response generalization is expected, since changes in other measured behaviors concomitant with changes in the experimentally treated behavior will prevent demonstration of a functional relation between the experimental variable and behavior change. Since the nature and extent of response generalization in token economies has not been adequately investigated, Kazdin and Bootzin are tentative in their suggested use of the above multiple-baseline design in token economy evaluation.

The difficulties encountered with the various experimental designs used to date in token economy research suggest the need for exploring alternative designs. The author is currently investigating the usefulness of the multiple-baseline design in evaluating a ward-wide token economy by dividing target behaviors into four main categories and introducing the token economy in stages, incorporating each category in sequential fashion. Preliminary data indicate that the multiple-baseline design can provide a powerful and practical method for token economy evaluation.

In the next section, existing token economy evaluation studies will be reviewed and critically examined. Emphasis will be placed on the importance

of (1) clearly specified treatment procedures, (2) adequate dependent measures, and (3) demonstration of a causal relationship between treatment and behavior change.

Review and Critique of Token Economy Evaluation Studies

Pre-Experimental Designs

Single-group studies. The pre-experimental single-group studies to be reviewed used pre-post or repeated measurement in a single group of subjects. Such designs may be thought of as a combination of case studies with a number of replications occurring simultaneously. Steffy et al., (1969) monitored several target behaviors and administered the Psychotic Reaction Profile (Lorr, O'Connor, & Stafford, 1960) and Minimal Social Behavior Scale (Farina, Arenberg, & Guskin, 1957) prior to, during, and one year after implementation of a token economy. The results indicated improvement on target behaviors as well as in global individual functioning. Operant procedures were described in detail, but behavioral criteria were generally not specified in objective terms nor was observer reliability data provided. Since the program involved token reinforcement on a very limited scale, caution must be exercised in linking global changes to the treatment procedures described.

McReynolds and Coleman (1972) summarized in detail the token reinforcement procedures which were applied in sequential fashion to target behaviors during the development of a token program. They reported increased percentages of patients engaging in reinforced behaviors as the program progressed and found an increased in discharge and transfer rates over baseline. Measurement methods and observation criteria for determining occurrence of target behaviors were not described, and no measurement of individual functioning or general ward behavior was reported.

The experimental designs employed by Steffy et al. (1969) and McReynolds and Coleman (1972) contained no controls for simultaneous extra-treatment events, maturation (effects of time), or instrumentation changes, and thus precluded demonstration of cause-effect relationships between treatment procedures and specific or global behavior changes.

Static-group comparisons. In a typical static-group comparison, existing groups are compared after a treatment condition is applied to one of the groups. Birky et al. (1971) reported more patients discharged (the difference was not significant) from a token economy than from two existing control wards. Treatment procedures were summarized, but not in detail. No systematic measurement of target behavior rates, individual functioning, or general ward behavior was reported. Since the type of experimental design used by Birky et al. (1970) provided neither pretesting nor assurance of group equivalence prior to treatment, a cause-effect relationship could not be established as a result.

Quasi-Experimental Designs

Time-series experiments. In a time-series experiment, subjects serve as their own controls. Periodic assessment is conducted prior to and after the introduction of a treatment condition. Treatment effects are suggested if a discontinuity occurs in the measurements recorded in the periodic assessment process. Atthowe and Krasner (1968) provided a general, over-all description of an extensive token economy and reported time-series

data which showed improvement from baseline to experimental conditions for several target behaviors. At one point, a slight increase over the average level of activity attendance occurred when the value of tokens for activity attendance was temporarily raised, but the increase in attendance was not dramatic and only slightly suggestive of a functional relationship. Discharges and transfers increased in frequency over the baseline period. A significant increase in social responsiveness at weekly meetings was measured by the Palo Alto Group Psychotherapy Scale (Finney, 1954). No systematic measurement of general ward behavior was reported; much of the information provided in the study was anecdotal in nature.

Gorham et al. (1970) made use of an existing quarterly rating system by nurses which involved rating patients on the Perry Point Nursing Evaluation Scale (VA Forms 10-47, 48, 49-641) at quarterly intervals. After 9 months the patients participating in a token program showed significant improvement on 8 of the 24 scale dimensions when compared to baseline performance. A substantial increase in problem behaviors appeared on the first quarterly rating after program implementation and was attributed by the authors to the new contingencies and loss of free privileges. However, possible effects of regrouping (De Vries, 1968) should also be considered since patients were selected and moved to the token economy ward during the interval preceding the measured increases. The token economy was described in summary form, with no treatment procedures specified in detail. No frequency data were provided for target behaviors or general ward behavior. The authors suggested that some of the changes in evaluation scale ratings may have reflected changes in the nurses' subjective perception of the patients as the program progressed, rather than actual changes in patient behavior. This possibility highlights the importance of controlling for instrument errors in time-series designs. In addition to possible confounding due to instrumentation changes, the time-series designs used by Atthowe and Krasner, and Gorham et al., provide no control for effects of simultaneous extra-treatment (historical) events and thus do not allow demonstration of causality (Campbell & Stanley, 1963).

Nonequivalent control group studies. In nonequivalent control group studies, experimental and control groups are given both a pretest and post-test. Assurance of group equivalence by means of random assignment is not provided. Instead, groups are developed by matching subjects on pre-determined dimensions, or existing groups are used. Heap et al. (1970) reported more patients discharged and transferred to better wards from a token economy than from a control ward after six months of treatment. The initial groups were not identical, however; twenty of the most problematic patients from the initial sample of 90 were assigned directly to the token program, while the remainder were randomly assigned, resulting in 45 per group. The token economy ward included individualized token reinforcement, ward government, and attitude therapies combined, which precluded an analysis of treatment effects attributable to the token economy alone. Time-series measurement revealed a gradual increase in performance of target behaviors reinforced with verbal praise and tokens, but no causal relationship was demonstrated. No systematic measurement of individual functioning or general ward behavior was reported. Follow-up data after 35 months revealed a substantially higher percentage of patients discharged, furloughed, and remaining in the community than the percentages for the remainder of the hospital. However, after the first six months of the project, patients were

assigned to the token economy as vacancies occurred, and it is quite possible that potentially dischargeable patients from other areas of the hospital were channeled through the token economy, which may have been better equipped for community placement and follow-up. Thus the token economy may have become a center for discharge, and not necessarily an effective treatment program.

Gripp and Magaro (1971) described in detail an extensive, well-staffed token economy and reported positive changes in relation to control wards using the Nurse's Observation Scale for Inpatient Evaluation (Honigfeld & Klett, 1965), Psychotic Reaction Profile (Lorr et al., 1960), Minimal Social Behavior Scale (Farina et al., 1957), Elgin Behavior Rating Scale (Wittman & Hills, 1947) and Ward Atmosphere Scale (Moss & Houts, 1968) as measures. However, most pretesting was done before the 45 patients meeting selection criteria were regrouped on the token economy ward. Therefore, reported changes could have been related to the effects of transfer and regrouping (DeVries, 1968). In addition, the authors indicated that medication for 80% of the token economy patients was discontinued shortly after program initiation, thus introducing another confounding factor. No measurement of specific target behaviors was reported.

A comparison of patient progress on a token economy with that of a matched control group on a physically identical ward was conducted by Shean and Ziedberg (1971). The groups consisted primarily of original ward populations with some transfers to achieve matching on age, diagnosis, and chronicity. The authors summarized the treatment program but did not specify procedural details. Token economy patients demonstrated pre-post improvements significantly greater than those of the control group on key tapping rates, Motility-Affect-Cooperation-Communication scale (Ellsworth, 1962) scores, and attendance at gym class and occupational therapy sessions. Outcome data also indicated that performance of housekeeping duties and time spent out of the hospital were greater, and medication levels lower for token economy than for control patients. However, baseline data on these latter variables were not provided, so it is impossible to determine whether the reported figures represent changes, existing group differences, or a combination of both. No systematic measurement of general ward behavior was described.

The studies by Heap et al. (1970), Gripp and Magaro (1971), and Shean and Ziedberg (1971) used designs in which control groups were established in a manner other than by random assignment, either by using existing groups or by matching patients on predetermined dimensions. It is difficult to determine a priori what important dimensions should be used for matching (Campbell & Stanley, 1963) and matching on a few variables such as age, diagnosis, and length of hospitalization may easily mask important qualitative differences in patient level of functioning and prognosis. Campbell and Stanley (1963) emphasized that randomization is essential for establishing group equivalence, and that matching may be a useful adjunct to randomization but not a substitute for it. Thus, in the above designs no definite cause-effect relationships could have been experimentally demonstrated even if other methodological flaws had not been present. However, to the extent that initial group equivalence is likely in such designs, results may be yielded which are suggestive of causal relationships.

Equivalent time-samples designs. In equivalent time-samples designs, periods of treatment and no treatment are presented in alternating fashion to allow investigation of treatment effects. Ayllon and Azrin (1965) conducted a series of six ABA own-control experiments to demonstrate functional

control of token reinforcement over both on and off-ward job performance and preference. Winkler (1970) and Kazdin and Bootzin (1972) have aptly criticized several of the experiments on grounds that verbal instructions and not contingent token reinforcement may have controlled work performance. Subjects were told that a non-contingent reinforcement phase would be a "vacation with pay" (Ayllon & Azrin, 1965) and thus may have perceived work as being inappropriate during that period. Nevertheless, control by the overall treatment package (verbal instructions with contingent token reinforcement) over job performance and control of specific token reinforcement procedures on job selection, were demonstrated by the experiments. Treatment procedures were specified in detail and dependent measures were clearly described. However, since the authors presented only job performance and job preference findings and did not report data regarding other important target behaviors, individual functioning levels or general ward behavior, this study cannot be considered an overall evaluation of a token economy program.

Lloyd and Garlington (1968) alternated contingent and noncontingent token reinforcement in an ABAB design to assess effects on performance of various self-care behaviors. Performance increased, though not dramatically, when contingent tokens were introduced, and decreased only slightly during the withdrawal condition. Treatment procedures were not described in detail. Measurement methods for determining occurrence of target behaviors were not reported, and no measurement of individual functioning or general ward behavior was discussed. The lack of substantial decrease in performance during reversal supports Kazdin and Bootzin's (1972) observation that the effects of token reinforcement on certain behaviors may not be transient, and therefore may not be appropriate for investigation through withdrawal designs.

Winkler (1970) evaluated the effects of both positive and negative response consequences on target behaviors within a token program. In several cases, time-series measurement revealed behavior changes in the direction desired, though no evidence of causality was provided by the time-series analysis. Since improvement was observed in some behaviors within the token economy for which reinforcement was not directly programmed, an ABA design was used to investigate the control of contingency procedures over four target behaviors. Functional control of the operant techniques over frequencies of shoe cleaning, performance of a work task, loud noise, and violence, was demonstrated. Observer reliability data or behavioral definitions for noise and violence were not provided, although systematic measurement procedures were described for several other target behaviors. Treatment procedures used to affect changes in specific response rates were described in detail. No systematic measurement of individual functioning levels or general ward behavior was reported.

The withdrawal design as employed by Ayllon and Azrin (1965), Lloyd and Garlington (1968) and Winkler (1970), allows demonstration of causal relationships between treatment procedures and behavior change (Campbell & Stanley, 1963; Paul, 1969). However, the above investigators applied the design to a few selected behaviors only, and thus did not establish the effects of operant procedures on other target behaviors within their programs. In addition, no assessment of generalized therapeutic effects over and above frequency changes in directly reinforced behaviors was provided in the above studies.

True Experimental Designs

Control group designs with random assignment. In this type of design, random assignment of subjects to experimental and control groups is conducted to achieve group equivalence prior to treatment. Pretests and posttests, posttests only, or repeated measures are used to evaluate treatment effects. Schaefer and Martin (1966) randomly assigned patients on a token economy ward to two groups and administered a supply of free tokens each morning to one group, while the other group received tokens contingent upon desired behavior. The authors provide a general summary of the treatment procedures, which included extensive use of individualized programming. A time-sampling procedure was conducted in which observed behavior of individual subjects was coded as mutually-exclusive, concomitant, or idiosyncratic. Each observation necessarily contained a mutually-exclusive behavior (sitting, standing, walking, running, lying down) and also typically included one of the concomitant (e.g., talking, smoking) or idiosyncratic (e.g., rocking pacing) behaviors. Instances in which only a mutually-exclusive behavior (i.e., sitting, standing, walking, running, lying down) was observed were rated as indicating inactivity or "apathy", and were totaled for each individual to obtain an "apathy" score. After two months on the program the token economy patients showed a significant decrease in apathy when compared with baseline and control levels. Although this study was one of the first token economy evaluations reported, it represents a rare demonstration of a cause-effect relationship between a total token economy and change on an objective, clearly-specified dependent measure. The measure used, however, did not provide information regarding the types of behaviors that increased as a result of treatment. Increases in any of the concomitant or idiosyncratic behaviors, whether appropriate or inappropriate, could have resulted in lower apathy scores. Since no data were presented regarding specific target behaviors, individual functioning levels, or time-sampled ward behaviors other than "apathy", no clear description of treatment effects was provided.

In a study reported by Henderson and Scoles (1970), subjects drawn from an independent receiving facility were randomly assigned to a state hospital, city hospital, or community-based token program. Subsequent data evaluation revealed group populations of 10, 9, and 21 respectively, but no explanation was provided for the large number reported in the token economy group. Outcome data after 549 days revealed that token economy patients, while requiring a longer initial hospitalization period, stayed in the community longer following release, remained employed longer, and required less rehospitalization than control patients. Since the token economy was combined with a community treatment approach involving substantial use of community volunteers and programmed activities within the community, caution must be exercised in attributing any reported therapeutic effects to operant components of the program. In addition to the above experimental-control comparisons, the authors analyzed several classes of behavior within the token program using time-series and reversal designs. Time-series data revealed steady increases in reinforced behaviors but behavioral decline under withdrawal conditions was not clear-cut. Neither the specific reinforcement techniques nor the procedures for defining and measuring behavioral change were reported. No assessment of individual functioning levels or general ward behavior was discussed.

Maley et al. (1973) randomly assigned patients to experimental and control groups. The control patients remained on their original units, while the experimental patients were assigned to a token economy ward. The

authors provided a general description of the token program, which included individualized token reinforcement and group treatment strategies. After approximately six months, the token economy patients demonstrated superior performance on a variety of measures of individual functioning when compared with controls. The assessment procedures, which included standardized interview tasks and objectively rated video recordings of interview behavior, were described in detail and observer reliability data provided. Since no pretest was conducted after arrival on the token economy ward, no control was provided for possible effects of transfer and regrouping (De Vries, 1968). Thus posttest differences cannot appropriately be attributed to treatment procedures alone. No data regarding specific target responses or general ward behavior were reported. The authors proposed a replication of the above study which would include pretests as well as posttests.

The control group designs with random assignment used by Schaefer and Martin (1966), Henderson and Scoles (1970), and Maley et al. (1973) generally allow demonstration of causal relationships. However, the presence of confounding factors or inadequate measurement procedures in the above studies resulted in less-than-conclusive findings.

Conclusion

No study to date has demonstrated a cause-effect relationship between clearly specified treatment procedures and positive changes in target behavior rates, individual functioning levels, and general ward behavior. Of the studies reviewed, only the report by Ayllon and Azrin (1965) specified treatment procedures in sufficient detail to allow replication; no studies investigated all three of the domains suggested for measurement of treatment effectiveness; and only the studies by Ayllon and Azrin (1965) Schaefer and Martin (1966), and Winkler (1970) clearly established causal relationships between treatment procedures and behavior changes.

Paul (1969) proposed a tactical sequence for systematic development and evaluation of behavior modification techniques, as follows: (a) laboratory development of treatment techniques, (b) trial and modification through case studies and single-group experiments, (c) evaluation of effectiveness in control group studies, and (d) analysis of treatment components and comparisons with other techniques using factorial experiments. It is clear that token economy development and evaluation have progressed through levels (a) and (b). It is not clear that the token economy approach has achieved a level of demonstrated effectiveness to satisfy the requirements of level (c).

The preceding critical analysis focused on methodological aspects of the studies reviewed, and therefore did not reveal the wide variety of innovative applications of operant principles described in many token economy studies. The task of developing a token economy appears to involve a number of formidable theoretical, tactical, and practical problems which may easily impinge on the maintenance of scientific rigor. The complex token economy within a hospital setting is not easily evaluated (Gripp & Magaro, 1974; Milby, 1975) and the methodological guidelines discussed herein are suggested as objectives rather than attempts to detract from the substantial technological advances evident in the token economy literature.

References

Agras, W. S. (Ed.), Behavior modification: Principles and clinical applications. Boston: Little, Brown and Company, 1972.

Aitchison, R. A., & Green, D. R. A token reinforcement system for large wards of institutionalized adolescents. Behavior Research and Therapy, 1974, 12, 181-190.

Allen, D., & Magaro, P. A. Measures of change in token-economy programs. Behaviour Research and Therapy, 1971, 9, 311-318.

Atthowe, J. M., & Krasner, L. Preliminary report on the application of contingent reinforcement procedures (token economy) on a "chronic" psychiatric ward. Journal of Abnormal Psychology, 1968, 73, 37-43.

Ayllon, T., & Azrin, N. H. The measurement and reinforcement of behavior of psychotics. Journal of the Experimental Analysis of Behavior, 1965, 8, 357-383.

Baer, D. M., Wolf, M. M., & Risley, T. R. Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Bandura, A. Principles of behavior modification. New York: Holt, Rinehart and Winston, 1969.

Bennett, P. S., & Maley, R. F. Modification of interactive behaviors in chronic mental patients. Journal of Applied Behavior Analysis, 1973, 6, 609-621.

Birky, H. J., Chambliss, J. E., & Wasden, R. A. Comparison of residents discharged from a token economy and two traditional psychiatric programs, Behavior Therapy, 1971, 2, 46-51.

Boren, J. J., & Colman, A. D. Some experiments on reinforcement principles within a psychiatric ward for delinquent soldiers. Journal of Applied Behavior Analysis, 1970, 3, 29-37.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, Pp. 171-246.

Carlson, C. G., Hersen, M., & Eisler, R. M. Token economy programs in the treatment of hospitalized adult psychiatric patients. Journal of Nervous and Mental Disease, 1972, 155, 192-204.

DeVries, D. L. Effects of environmental change and of participation on the behavior of mental patients. Journal of Consulting and Clinical Psychology, 1968, 32, 532-536.

Ellsworth, R. B. The MACC behavioral adjustment scale, Form II. Beverly Hills: Western Psychological Services, 1962.

Farina, A., Arenberg, D., and Guskin, S. A scale for measuring minimal social behavior. Journal of Consulting Psychology, 1957, 21, 265-268.

Finney, B. S. A scale to measure interpersonal relationships in group psychotherapy. Group Psychotherapy, 1954, 7, 52-66.

Gorham, D. C., Green, L. W., Caldwell, L. R., & Bartlett, E. R. Effect of operant conditioning techniques on chronic schixophrenics, Psychological Reports, 1970, 27, 223-234.

Gripp, R. F., & Magaro, P. A. A token economy program evaluation with control ward comparisons. Behavior Research and Therapy, 1971, 9, 137-149.

Gripp, R. F., & Magaro, P. A. The token economy program in the psychiatric hospital: A review and analysis. Behaviour Research and Therapy, 1974, 12, 205-228.

Heap, R. F., Boblitt, W. E., Moore, C. H., & Hord, J. E. Behavior-milieu therapy with chronic neuropsychiatric patients. Journal of Abnormal Psychology, 1970, 76, 349-354.

Henderson, J. D., & Scoles, P. E. Conditioning techniques in a community-based operant environment for psychotic men. Behavior Therapy, 1970, 1, 245-251.

Hersen, M., Eisler, R. M., Smith, B. S., & Agras, W. S. A token reinforcement ward for young psychiatric patients. American Journal of Psychiatry, 1972, 129, 142-147.

Honigfeld, G., & Klett, C. J. The nurses observation scale for inpatient evaluation: A new scale for measuring improvement in chronic schizophrenia. Journal of Clinical Psychology, 1965, 21, 65-71.

Kazdin, A. E. The token economy: A review and evaluation. New York: Plenum Press, 1977.

Kazdin, A. E., & Bootzin, R. R. The token economy: An evaluative review. Journal of Applied Behavior Analysis, 5, 343-372.

Kiesler, D. J. Experimental designs in psychotherapy research. In A. E. Bergin and S. L. Garfield (Eds.), Handbook of psychotherapy and behavior change. New York: John Wiley and Sons, 1971.

Krasner, L., & Atthowe, J. M. The token economy as a rehabilitative procedure in a mental hospital setting. In H. C. Richard (Ed.), Behavioral intervention in human problems. New York: Pergamon Press, 1971, 311-333.

Lloyd, K. E., & Abel, L. Performance on a token economy psychiatric ward: A two year summary. Behavior Research and Therapy, 1970, 8, 1-9.

Lloyd, K. E., & Garlington, W. K. Weekly variations in performance on a token economy psychiatric ward. Behavior Research and Therapy, 1968, 6, 407-410.

Lorr, M., O'Connor, J. P., & Stafford, J. W. The psychotic reaction profile. Journal of Clinical Psychology, 1960, 16, 241-245.

Maley, R. F., Feldman, G. L., & Ruskin, R. S. Evaluation of patient improvement in a token economy treatment program. Journal of Abnormal Psychology, 1973, 82, 141-144.

Marks, J., Sonoda, B., & Schalock, R. Reinforcement vs. relationship therapy for schizophrenics. Journal of Abnormal Psychology, 1968, 73, 397-402.

McReynolds, W. T., & Coleman, J. Token economy: Patient and staff changes. Behavior Research and Therapy, 1972, 10, 29-34.

Milby, J. B. A review of token economy treatment programs for psychiatric inpatients. Hospital and Community Psychiatry, 1975, 26, 651-658.

Moos, R. H., & Houts, P. S. Assessment of social atmospheres of psychiatric wards. Journal of Abnormal Psychology, 1968, 73, 595-604.

Paul, G. L. Behavior modification research: Design and tactics. In Franks, C. M. (Ed.), Behavior therapy: Appraisal and status, New York: McGraw-Hill, 1969, 29-62.

Schaefer, H. H., & Martin, P. L. Behavioral therapy for "apathy" of hospitalized schizophrenics. Psychological Reports, 1966, 19, 1147-1158.

Shean, G. D., & Zeidberg, Z. Token reinforcement therapy: A comparison of matched groups. Journal of Behavior Therapy and Experimental Psychiatry, 1971, 2, 94-105.

Sidman, M. Tactics of scientific research. New York: Basic Books, 1960.

Steffy, R. A., Hart, J., Craw, M., Torney, D., & Marlett, N. Operant behavior modification techniques applied to a ward of severely regressed and aggressive patients. Canadian Psychiatric Association Journal, 1969, 14, 59-67.

Tenbrunsel, T. W. A group coupon economy program in treating mental illness: An Experiment in social innovation. Unpublished doctoral dissertation St. Louis University, 1969.

Winkler, R. C. Management of chronic psychiatric patients by a token reinforcement system. Journal of Applied Behavior Analysis, 1970, 3, 47-55.

Wittman, P., & Hills, O. Elgin behavior rating scale. Elgin Papers, Elgin State Hospital, Elgin Illinois, 1947 (mimeograph copy).